Update on the Demographic Profile and Demographic and Housing Characteristics File (DHC)

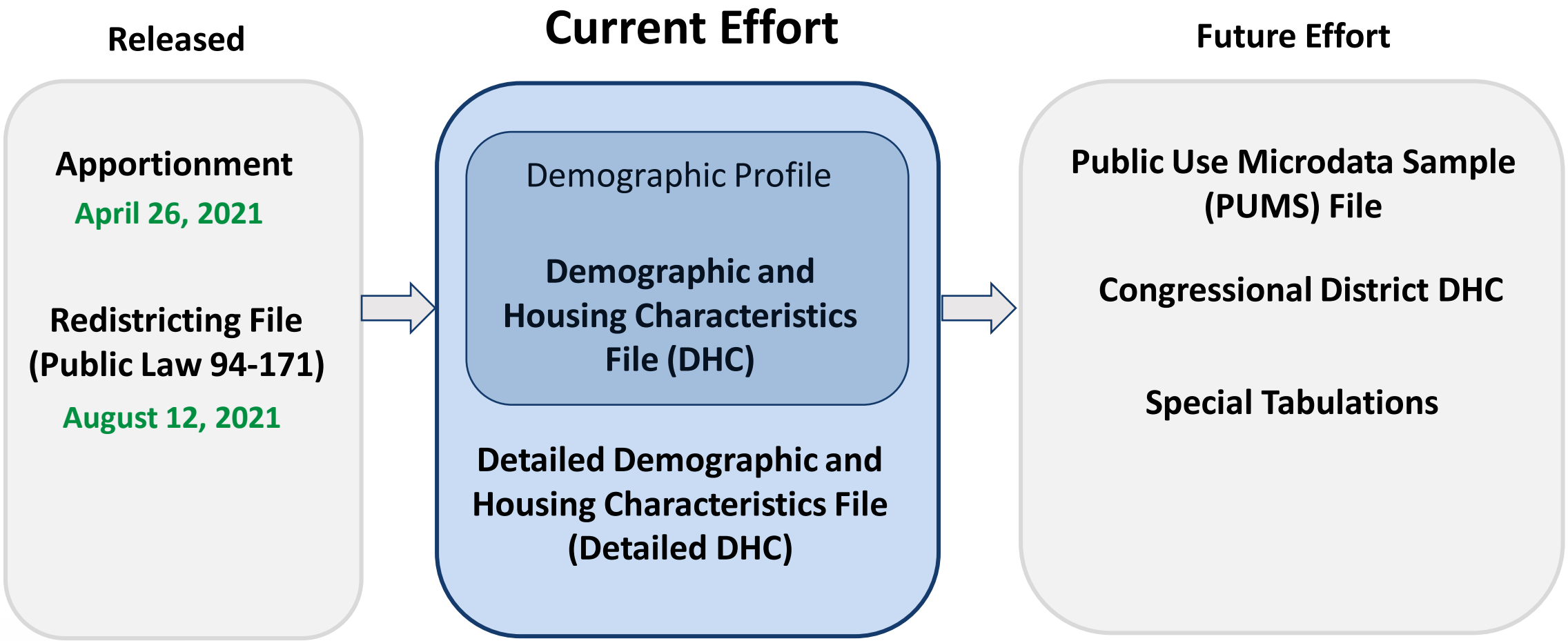
Jason Devine
Population Division

Matthew Spence
Population Division

Census Scientific Advisory Committee 2022 Spring Virtual Meeting
March 17-18, 2022

United States® Census Bureau

2020 Census Data Products



United States® Census Bureau

2020 Census Data Products: DHC and Demographic Profile

Demographic and Housing Characteristics File (DHC)

Includes many of the demographic and housing tables previously included in the 2010 Census Summary File 1. The DHC will include counts of households or families by type but not counts of people by those types. It will however, include counts of people by relationship categories. Geography varies with many tables provided at the Census Block level.

Demographic Profile

Provides critical demographic and housing characteristics about local communities from selected DHC tables in a streamlined, easy to use format. In previous decades, this product was released as soon after the release of the Redistricting product as possible. With the implementation of the new disclosure avoidance methodology the production of the Demographic Profile tables follows the same timeline as the DHC. The current plans include for the release of the Demographic Profile at the same time as the DHC.

Understanding the DHC: Background

- The Census Bureau is implementing new disclosure avoidance methods based on differential privacy to protect against an increased disclosure risk because of an increased availability of external data and advances in computer science
- The timeline for the development and release of the Demographic and Housing Characteristics File (DHC) was impacted by splitting the development of the Disclosure Avoidance System (DAS) for the Redistricting data from the DAS for the DHC and Detailed DHC
- Development work on the DAS was further split between the DHC and Detailed DHC
- Use cases from comments and data accuracy are being considered in decisions about the demographic and geographic detail included in the DHC and Detailed DHC
- Consistency between the Redistricting and DHC data continues to be included in the design of the DAS

Understanding the DHC: Topics Included in the 2020 Census:

- Relationship to Householder
- Sex
- Age
- Hispanic or Latino Origin
- Race
- Housing Tenure (owned, rented, occupied without payment)
- Vacancy Status

United States® Census Bureau

Understanding the DHC: Topics Included in the DHC

Counts of People by:

- Sex by single year-of-age
- Hispanic or Latino origin of householder by race of householder
- GQ population by sex by age
- Relationship by age for population under 18 years
- Household type by relationship and presence of people of specific ages

Counts of Households/Families by:

- Multigenerational households
- Family type by presence of children
- Tenure by household size
- Tenure by household type by age of householder
- Vacancy Status

Lowest level of geography: Various with many tables proposed at Census Block

Understanding the DHC: Determining the Contents

- The initial proposal for the design and contents of the DHC was developed and shared in 2019 based on subject matter expert knowledge and feedback from a Federal Register Notice
- The 2010 Census Summary File 1 (SF1) was the starting point for the design of the DHC with the DHC including many of the tables included in SF1
- The contents of the DHC were reviewed and updates proposed based on the feedback received documenting uses of the data. The cost for privacy and the accuracy of the new tables must be considered before they are added to the DHC
- We are looking at options for sharing information from the comments we received
- Changes to the contents of the DHC will be reflected in the second DHC demonstration product

Comment Summary

- **Over 400** comments received on the DHC and Detailed DHC throughout the October and December 2021 comment periods
- Many of the comments provided specific use cases and data needs
- Most uses that require basic demographic and housing information will be broadly met by the tables currently proposed for the DHC
- Comments documented uses of household and family type tables at lower levels of geography mainly for larger cities for housing and resource allocation
- Those who produce estimates and projections for planning purposes use age and sex detail by race and ethnicity for lower levels of geography than currently planned for the DHC
- Group quarters data by race and ethnicity are also used at lower levels of geography for projections, policy, and planning than currently planned for the DHC
- Analysis is being conducted to determine if tables can be added to the DHC to better meet the needs of the uses documented in the comments

Understanding the DHC: Protecting from Disclosure

- Data from the Census Edited File (CEF) are coded to create only the variables needed to generate the tables in the DHC and then put into Person and Unit (Unit) Microdata files
- Variables in the DHC Person and Unit Microdata files are protected using the TopDown algorithm (TDA)
  - Redistricting data is used as input with DHC tables matching the Redistricting tables exactly
- Data undergo post processing, which ensures certain consistencies
  - For example, total counts will be consistent within and across tables
  - Comparisons between Person and Unit file tables may reveal inconsistencies, for example, tables with counts of people compared to tables with counts of households
- Census creates tables for dissemination by tabulating the protected DHC Person and Unit Microdata files

Understanding the DHC: Experiments to Improve Accuracy

- A disclosure avoidance system (DAS) was developed to support the DHC using default settings
- A set of accuracy targets or ideal levels of error for variables and crosstabulations was developed
- Experimental runs using 2010 Census data allow us to assess the resulting privacy and accuracy
- A re-identification study is used to assess the risk of re-identification
- Leading experimental run along with the results for the accuracy analysis and re-identification study are discussed with DSEP to obtain approval to release as the demonstration product
- The demonstration product is made available to the public for their analysis with their comments helping guide work on improvements to the DAS

Understanding the DHC: The Demonstration Product

- A DHC demonstration product for tables sourced from the **Person File** in the form of Summary Files was released for public analysis on March 16
- Accuracy metrics for the tables sourced from both the **Person and Unit Files** were also released on March 16
- A DHC demonstration product for tables sourced from the **Unit File** is being prepared for release
- Two rounds of demonstration products are planned for the DHC (Winter 2022, Spring 2022)
- The public will have 30 days to review and comment on the **Person File** and 30 days to review and comment on the **Unit File** demonstration products
- This is the most extensive set of demonstration data to be released using the new disclosure avoidance methods (114 Person File tables, 119 Unit File tables, 66 different geographic levels)
- A webinar is planned for March 22 to inform the public about the demonstration products

Understanding the DHC: Internal Analysis of the DHC Person and Unit Demonstration Product

What We Looked For

- Odd, implausible, or impossible results
- Measures of average difference that exceed standards (numeric or percent)
- Differences that change patterns or that would change the story or headline
- Inconsistencies from applying disclosure avoidance to the Person and Unit files independently

Key Findings from the Person and Unit File

- In some cases, the GQ Population was assigned a Detailed GQ type not present within the state
- Some small groups have small numeric but large percent differences with a positive bias
- Larger differences for some ages such as 18 to 24-year-olds (GQ population) and centenarians
- Lower accuracy for counts of people by some relationship to the householder categories
- Bias towards higher counts of households with children present
- Lower accuracy for same-sex married and same-sex unmarried couple households

Evaluating DHC Demonstration Data

Detailed Summary Metrics

- Create measures of **accuracy**, **bias**, and **outliers** based on tabulated quantities at various geographic levels (e.g., Total Population at the county level, Asian Alone at the tract level) for a DAS run to published (i.e., swapped) 2010 tabulations

Tabulations / Maps

- Directly compare values using tables or maps to illustrative substantive accuracy/error

United States® Census Bureau

Mean Absolute Error for Sex by Age Groups – Counties

	Demonstration Data #1 (2019)	Demonstration Data #2 (2020)	Demonstration Data #3 (2022) ρ = 3.3	DAS Run ρ = 6.6	High Swap
Total					
0 to 17 years	46.43	42.95	**9.84**	9.84	256.41
18 to 64 years	99.17	231.70	**12.83**	12.57	494.16
65 years and over	103.04	230.85	**12.66**	12.22	431.37
Male					
0 to 17 years	65.52	92.39	**7.59**	7.33	137.63
18 to 64 years	90.78	195.33	**10.29**	9.98	265.76
65 years and over	74.38	137.97	**7.31**	6.85	213.32
Female					
0 to 17 years	64.36	93.18	**7.47**	7.31	133.00
18 to 64 years	93.85	181.53	**8.03**	7.61	273.48
65 years and over	73.49	147.95	**7.24**	6.83	225.25

Mean Absolute Error for 5 Year Age Groups – Counties

	Demonstration Data #1 (2019)	Demonstration Data #2 (2020)	Demonstration Data #3 (2022) ρ = 3.3	DAS Run ρ = 6.6	High Swap
Under 5 years	96.50	83.67	**7.75**	6.70	107.33
5 to 9 years	111.47	81.73	**7.42**	6.32	90.20
10 to 14 years	110.19	81.50	**7.36**	6.55	96.83
15 to 19 years	100.38	92.31	**12.03**	10.59	94.08
20 to 24 years	100.12	106.48	**18.20**	15.00	270.86
25 to 29 years	107.86	95.39	**16.06**	12.98	235.94
30 to 34 years	114.02	89.38	**12.53**	10.23	175.05
35 to 39 years	97.79	87.08	**12.17**	9.98	133.54
40 to 44 years	97.65	84.14	**12.07**	9.44	117.71
45 to 49 years	116.99	87.00	**11.95**	9.48	120.84
50 to 54 years	110.95	84.56	**11.17**	9.43	112.05
55 to 59 years	97.19	82.07	**10.93**	9.09	109.40

United States® Census Bureau

Mean Absolute Error for 5 Year Age Groups – Counties

	Demonstration Data #1 (2019)	Demonstration Data #2 (2020)	Demonstration Data #3 (2022) ρ = 3.3	DAS Run ρ = 6.6	High Swap
60 to 64 years	99.92	81.74	**6.76**	5.59	125.97
65 to 69 years	99.45	76.22	**7.02**	6.03	134.60
70 to 74 years	99.40	76.12	**9.69**	8.05	113.74
75 to 79 years	85.14	72.49	**9.43**	7.52	90.91
80 to 84 years	80.09	69.43	**8.87**	7.20	69.87
85 to 89 years	78.29	65.74	**23.63**	18.58	43.12
90 to 94 years	65.03	53.44	**16.83**	13.65	18.04
95 to 99 years	39.15	35.26	**14.71**	11.71	5.64
100 to 104 years	10.77	12.86	**6.46**	5.56	1.31
105 to 109 years	3.71	2.80	**2.00**	1.81	0.23
110 to 115 years	4.27	2.43	**2.24**	1.78	0.04

Mean Absolute Error for Household Type – Counties

	Demonstration Data #1 (2019)	Demonstration Data #2 (2020) *	Demonstration Data #3 (2022)
Family households			
Married couple family	270.84	N/A	**111.46**
Other family			
Male householder, no spouse present	176.15	N/A	**37.70**
Female householder, no spouse present	230.08	N/A	**45.94**
Nonfamily households			
Householder living alone	65.92	N/A	**12.17**
Householder not living alone	163.83	N/A	**53.89**

* The second demonstration data (dated 2020-05-27) did not include housing unit data

Questions for CSAC

- What questions or concerns do you have about the DHC demonstration products?
- What questions or concerns do you have about the content planned for the DHC?
- Do you have any questions or concerns about any of the information presented today?

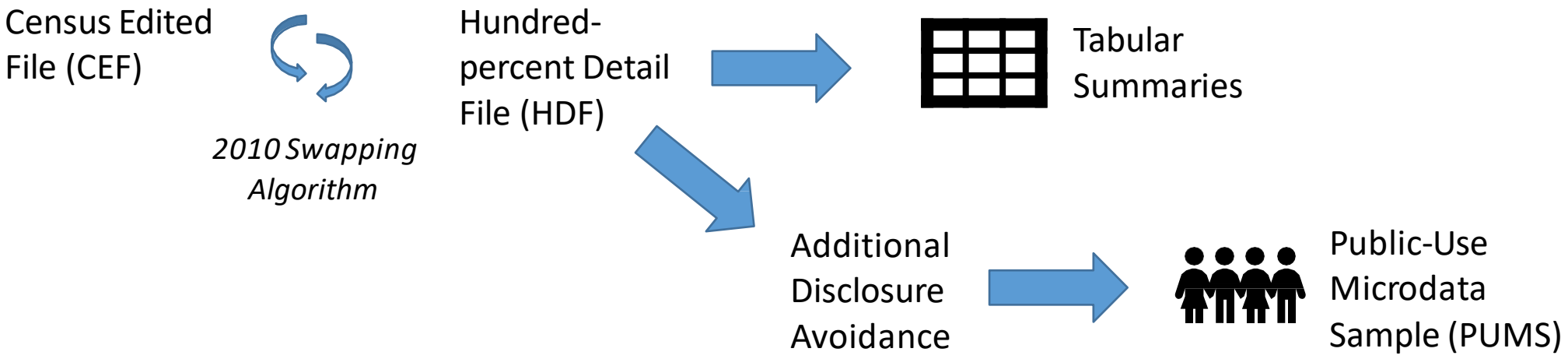
Reconstruction and Re-identification of the Demographic and Housing Characteristics File (DHC)

Michael Hawes
Senior Advisor for Data Access and Privacy
Research and Methodology

Presentation to the Census Scientific Advisory Committee
March 17-18, 2022

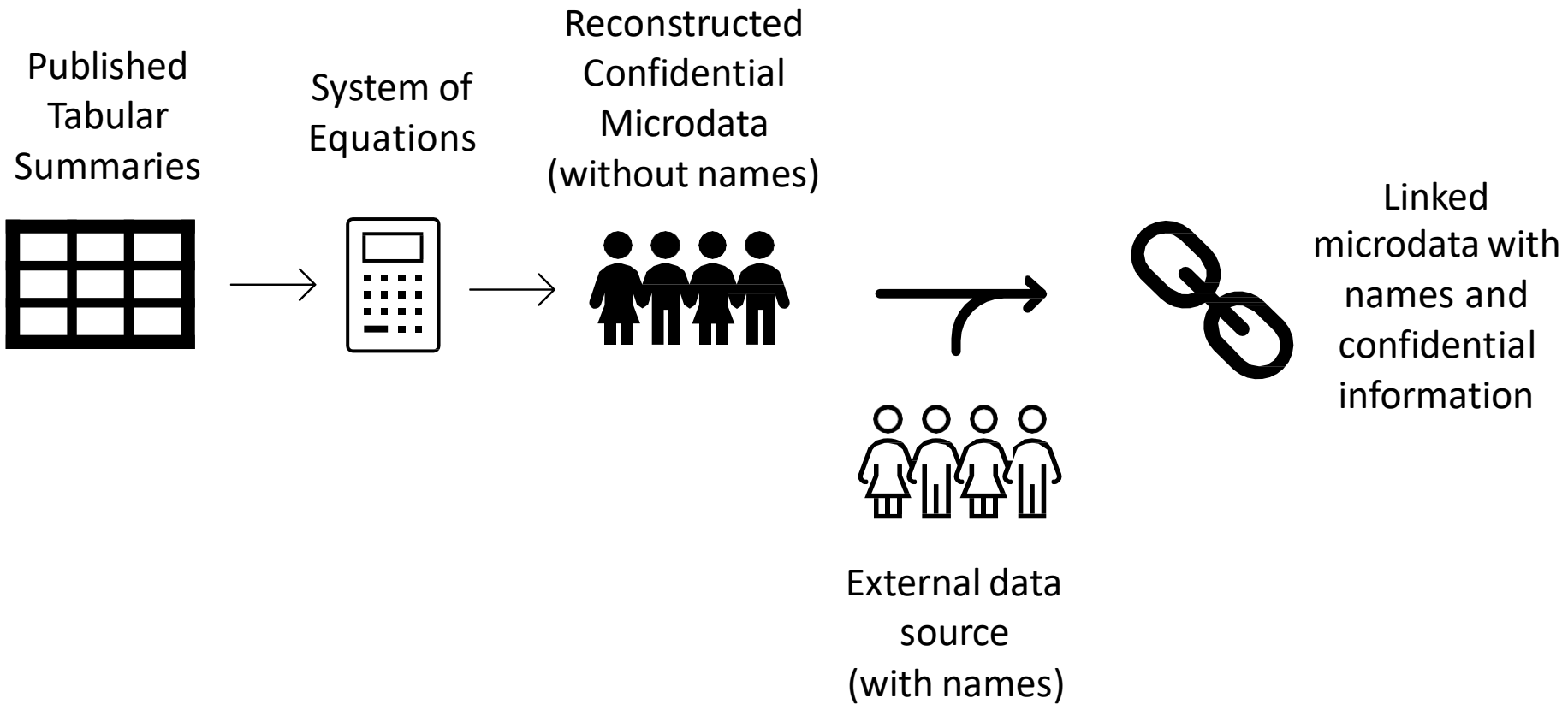
United States® Census Bureau

Disclosure avoidance for the 2010 Census



United States® Census Bureau

What is Reconstruction-abetted Re-identification?



United States® Census Bureau

What Census data can an attacker use?

Anything published from the 2010 Census!

Our simulated attack used only a small subset:

- P001 (Total Population by Block)
- P006 (Total Races Tallied by Block)
- P007 (Hispanic or Latino Origin by Race by Block)
- P009 (Hispanic or Latino, and Not Hispanic or Latino by Race by Block)
- P011 (Hispanic or Latino, and Not Hispanic or Latino by Race for the Population 18 Years and Over by Block)
- P012 (Sex by Age by Block)
- P012A-I (Sex by Age by Block, iterated by Race)
- P014 (Sex by Single-year-of-age for the Population under 20 Years by Block)
- PCT012A-N (Sex by Single-year-of-age by Tract, iterated by Race)

United States® Census Bureau

What external files can an attacker use?

Any external files that contain name and address (or other unique identifiers) and pseudo-identifiers contained in the census data (e.g., sex and age)

Our simulated attack used a combination of 4 commercially available datasets. But there are higher quality data out there. This is a lower-bound analysis.

The impact of higher quality name and address data can be estimated by using the CEF as the external file. This is an upper-bound analysis.

United States® Census Bureau

Exact Age vs. Binned Age

The subset of tables we used for our simulated attack do not always provide precise age reconstruction. Thus, we present re-id statistics for Exact Age matches, and for Binned Age matches using the following age bins from the block-level 2010 Summary File 1 tables:

Single year of age from 0 - 21
22-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-61
62-64
65-66
67-69
70-74
75-79
80-84
85+

United States® Census Bureau

Agreement rates for reconstructed microdata

Percentage of reconstructed records that exactly agree with the CEF on location, sex, age (exact/binned), race, and ethnicity

Agreement Rates	Exact Age	Exact and Binned Age
Published 2010 Tables (swapping)	46.5	91.8
High Swapping Experiment	26.5	52.1
DAS Run (ρ=3.325) (DDP*)	15.7	33.1
DAS Run (ρ=6.65)	17.1	36.4

*DDP: The DAS Run with ρ=3.325 is the run used to generate the 2010 DHC Demonstration Data Product 2022-03-16.

How well accurately can an attacker re-identify the characteristics of specific individuals from the reconstructed records?

United States® Census Bureau

Defining the universe for analysis

Successfully re-identifying specific individuals requires more than just a match on location, sex, age, race, and ethnicity.

It also requires being able to link a name to that record.

Not all records in the CEF have unique Protected Identification Key (PIK) identifier within the block.*

To evaluate the success of our simulated attack, we define the universe (denominator) as the data-defined population (individuals with unique PIKs within the block).

*A PIK is the Census Bureau's individual record linkage identifier produced by the Person Identification Validation System (PVS), the production name and address linkage system. The vintage is the same as for the 2010 Census.

United States® Census Bureau

Definitions

Putative Re-identification Rate:

$$\frac{\text{\# ooo mmmoommr tttttt ttoammr ooo BBBBoorrBB,SSnSS, AAoorr(rrSSttrrtt,bbbooooomr)}}{\text{\# ooo mmmoommr wwbbtt uobbuur PPPP bbo bbBBorrBB}}$$

Confirmed Re-identification Rate:

$$\frac{\text{\# ooo mmmoommr tttttt ttoommr ooo PPPPP,BBBBoorrBB,SSnSS, AAoorr(rrSSttrrtt,bbbooooomr) RRttmr ttoorrEEttodbrbbtbEE}}{\text{\# ooo mmmoommr wwbbtt uobbuur PPPP bbo bbBBorrBB}}$$

Re-identification Precision Rate:

$$\frac{\text{\# ooo mmmoommr tttttt ttoommr ooo PPPPP,BBBBoorrBB,SSnSS, AAoorr(rrSSttrrtt,bbbooooomr) RRttmr ttoorrEEttodbrbbtbEE}}{\text{\# ooo mmmoommr tttttt ttoommr ooo BBBBoorrBB,SSnSS, AAoorr(rrSSttrrtt,bbbooooomr)}}$$

United States® Census Bureau

Re-identification statistics

(Exact and Binned Age)	Putative Rate	Confirmation Rate	Precision Rate
Published 2010 Tables (swapping) to Commercial	60.2	24.8	41.2
High Swapping Experiment to Commercial	56.2	17.2	30.7
DAS Run (ρ=3.325) to Commercial (DDP)	38.5	11.1	28.7
DAS Run (ρ=6.65) to Commercial	41.0	12.1	29.5
Published 2010 Tables (swapping) to CEF	97.0	75.5	77.8
High Swapping Experiment to CEF	75.4	46.6	61.8
DAS Run (ρ=3.325) to CEF (DDP)	24.6	14.0	56.9
DAS Run (ρ=6.65) to CEF	47.9	30.1	62.9

Re-identification of population uniques

Re-identification statistics for "population uniques" of the linking pseudo-identifiers (those who are unique within their block on sex, and either exact age [SAB] or binned age [SAbB])

	Putative Rate		Confirmation Rate		Precision Rate	
	SAB	SAbB	SAB	SAbB	SAB	SAbB
Published 2010 Tables (swapping) to Commercial	32.9	23.1	28.3	21.8	86.0	94.6
High Swapping Experiment to Commercial	26.6	17.6	18.2	12.7	68.5	72.4
DAS Run (ρ=3.325) to Commercial (DDP)	13.4	7.0	8.9	4.7	66.6	66.5
DAS Run (ρ=6.65) to Commercial	14.6	7.7	9.9	5.2	67.8	67.8
Published 2010 Tables (swapping) to CEF	95.0	93.2	84.2	87.2	88.6	93.6
High Swapping Experiment to CEF	69.2	64.0	46.7	44.5	67.5	69.6
DAS Run (ρ=3.325) to CEF (DDP)	30.3	22.1	19.6	13.9	64.6	62.9
DAS Run (ρ=6.65) to CEF	33.3	24.4	22.0	15.7	66.0	64.5

United States® Census Bureau

Re-identification of Population Uniques for Non-Modal Races

Re-identification statistics for "population uniques" of the linking pseudo-identifiers (those who are unique within their block on sex, and either exact age [SAB] or binned age [SAbB]) for individuals of the blocks' Non-Modal Races

Non-Modal Race	Putative Rate		Confirmation Rate		Precision Rate	
	SAB	SAbB	SAB	SAbB	SAB	SAbB
Published 2010 Tables (swapping) to Commercial	24.0	13.8	14.3	12.3	59.4	89.2
High Swapping Experiment to Commercial	20.6	11.5	5.0	3.5	24.4	30.6
DAS Run (ρ=3.325) to Commercial (DDP)	11.4	5.3	2.4	1.2	20.8	23.2
DAS Run (ρ=6.65) to Commercial	12.3	5.7	2.6	1.4	21.2	24.0
Published 2010 Tables (swapping) to CEF	90.6	86.2	60.4	70.2	66.7	81.4
High Swapping Experiment to CEF	71.6	65.5	20.0	21.9	27.9	33.4
DAS Run (ρ=3.325) to CEF (DDP)	34.7	25.9	7.8	6.2	22.3	24.0
DAS Run (ρ=6.65) to CEF	37.6	28.3	8.6	7.1	23.0	25.1

Re-identification of Population Uniques for Modal Races

Re-identification statistics for "population uniques" of the linking pseudo-identifiers (those who are unique within their block on sex, and either exact age [SAB] or binned age [SAbB]) for individuals of the blocks' Modal Races

Modal Race	Putative Rate		Confirmation Rate		Precision Rate	
	SAB	SAbB	SAB	SAbB	SAB	SAbB
Published 2010 Tables (swapping) to Commercial	35.1	25.3	31.8	24.2	90.5	95.3
High Swapping Experiment to Commercial	28.0	19.0	21.4	14.9	76.3	78.5
DAS Run (ρ=3.325) to Commercial (DDP)	13.8	7.4	10.5	5.5	75.8	73.9
DAS Run (ρ=6.65) to Commercial	15.2	8.2	11.7	6.1	76.9	75.2
Published 2010 Tables (swapping) to CEF	96.1	94.8	90.0	91.3	93.6	96.3
High Swapping Experiment to CEF	68.6	63.6	53.2	50.0	77.5	78.5
DAS Run (ρ=3.325) to CEF (DDP)	29.2	21.2	22.4	15.8	76.7	74.3
DAS Run (ρ=6.65) to CEF	32.3	23.5	25.2	17.8	78.1	75.8

United States® Census Bureau

Summary of Re-identification statistics

Difference in percentage points across experiments, relative to the Published 2010 Tables, in the putative rate and precision rates. All rates are to CEF.

	High Swap		DAS Run (ρ=3.325) (DDP)		DAS Run (ρ=6.65)	
Metric	Putative Rate	Precision Rate	Putative Rate	Precision Rate	Putative Rate	Precision Rate
National	-22.6	-16.0	-72.4	-20.9	-49.1	-14.9
SAB Uniques	-25.8	-21.1	-64.7	-24.0	-61.7	-22.6
SAB Non-Modal	-19.0	-38.8	-55.9	-44.4	-53.0	-43.7
SAB Modal	-27.5	-16.1	-66.9	-16.9	-63.8	-15.5

United States® Census Bureau